Learning Topographic Representations for Linearly Correlated Components

Learning Topographic Representations for Linearly Correlated Components

Hiroaki Sasaki[1], **Michael U. Gutmann**[2], **Hayaru Shouno**[1] **and Aapo Hyvärinen**[2]

[1]Dept of Information and Communication Engineering
The University of Electro-Communications
hsasaki@cc.uec.ac.jp, shouno@ice.uec.ac.jp

[2]Dept of Mathematics and Statistics, Dept of Comp Sci and HIIT
University of Helsinki
{michael.gutmann, aapo.hyvarinen}@helsinki.fi

Abstract

Recently, some variants of independent component analysis (ICA) have been proposed to estimate topographic representations. In these models, the assumptions of ICA are slightly relaxed: adjacent components are allowed to have higher order correlations while being linearly uncorrelated. In this paper, we propose a new statistical model for the estimation of topographic representations. In the proposed model, the estimated components are sparse and linearly correlated. To confirm the behavior of the model, we perform experiments on artificial data. In applications of the model to real data, we find emergence of a new kind of topographic representation for natural images and the outputs of simulated complex cells in the primary visual cortex.

1 Introduction

Independent component analysis (ICA) is a statistical model to estimate non-Gaussian components from observed data. The basic model is

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{s}, \tag{1}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^t$ is the data vector, $\boldsymbol{s} = (s_1, s_2, \ldots, s_d)^t$ is the vector of non-Gaussian, independent components, and $\mathbf{A}$ is the mixing matrix. By supposing that $\mathbf{A}$ is invertible, the only problem to solve is the estimation of $\mathbf{A}$. ICA has seen applications in various kinds of fields such as computational neuroscience [1] or EEG/MEG analysis [2].

In basic ICA, there are limitations for the estimation. Especially, the order of the estimated components is not defined. To define a meaningful ordering, various kinds of extensions have been proposed. One typical model is topographic ICA (TICA) [3]. The key idea in TICA is to slightly relax the assumption of ICA: only adjacent components have energy correlations, and distant components are as independent as possible. Thus, the ordering in TICA is defined using statistical dependencies, which means that the order carries meaningful information. For the case of natural images, adjacent basis vectors turn out to have similar properties and the whole topographic map unveils a global ordering. More recently, Osindero and colleagues [4] proposed another energy-based model which produces similar results as TICA. These extensions model the correlations between the variances of the components. Other work along this direction is [5].

In this paper, we propose an alternative statistical model to estimate a topographic representation. Unlike previous work, we define the topography (ordering) based on linear correlation. This linear

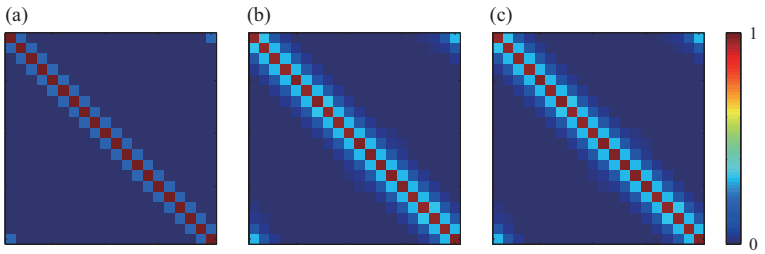

Figure 1: The covariance matrices of (a) sampled $\boldsymbol{s}$, estimated components (b) without and (c) with dynamic programming.

correlation can be seen in many practical situations. For example, consider natural images and the outputs of two co-linear Gabor filters with slightly different spatial positions along the preferred orientation, but with otherwise same parameters. When long edges or contours are often aligned with the preferred orientation, linear correlation in the Gabor outputs would be observed. Another example is coherent sources in EEG or MEG, which can be linearly correlated due to neuronal interactions [6].

In addition to constructing the new model, we propose an optimization algorithm which tends to avoid local maxima of the likelihood. Experiments on artificial and real data are performed to verify the utility of the model and the algorithm, and to investigate the difference to previous models.

This paper is organized as follows. First, the idea of a topography (ordering) is defined and motivated. Then, a probability density function (pdf) and the objective function are proposed in section 2. In section 3, numerical experiments on artificial data are performed. Then, we derive a new optimization method based on dynamic programming in an attempt to avoid local maxima. In section 4, we apply our model to natural images: we learn a topographic representation of both the raw images and the data after the application of a fixed complex cell stage. The connection to previous work and conclusions are given in section 5.

2 A new topographic model for correlated sparse components

The motivation for the topographic arrangement of the components based on statistical dependencies is as follows: First, it allows to easily visualize the interrelation between the components. Second, the topography which emerges for natural stimuli, for example natural images, may be related to cortical representations. As motivated in the last section, the statistical dependency which we use to define the topography is linear correlation,

$$E\{s_i s_{i+1}\} > 0. \tag{2}$$

To develop the model, we begin with the generative model

$$\boldsymbol{s} = \boldsymbol{u} \odot \boldsymbol{v}, \tag{3}$$

where $\odot$ denotes element-wise multiplication, $\boldsymbol{u}$ is a positive random vector and $\boldsymbol{v}$ is a multivariate Gaussian vector with mean zero. In Eq.(3), $\boldsymbol{u}$ and $\boldsymbol{v}$ are statistically independent. The key properties of this generative model are:

1. It generates super-Gaussian (sparse) components $\boldsymbol{s}$ [3].
2. It generates correlated sparse components $\boldsymbol{s}$ when the components in $\boldsymbol{u}$ are independent but the adjacent components in $\boldsymbol{v}$ are linearly correlated.

A similar model is used in TICA. Unlike here, however, the sources $\boldsymbol{s}$ in TICA are linearly uncorrelated and only have higher order correlations.

In principle, we could specify a prior for $\boldsymbol{u}$ and try to compute the marginal distribution of $\boldsymbol{s}$ (possibly by Monte Carlo Methods). However, we prefer here to simply modify the Laplace distribution, and

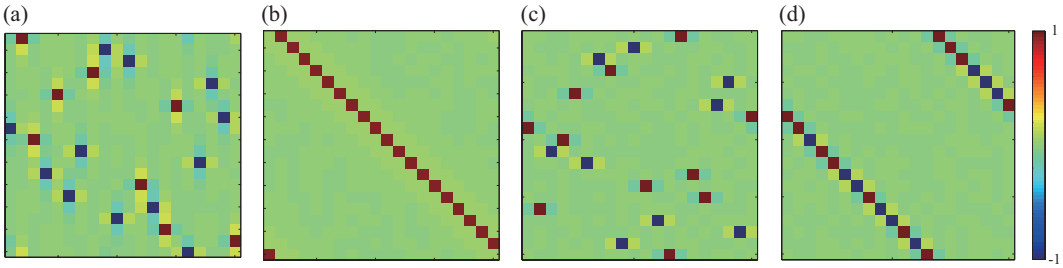

Figure 2: Matrices $\mathbf{P} = \mathbf{WVA}$. (a) and (b) are results from the proposed model without and with dynamic programming, respectively. (c) and (d) are results from TICA for the same artificial data. For (d), the ordering was performed manually.

use the distribution

$$p(\boldsymbol{s}) = \frac{1}{Z} \prod_{i=1}^{d} \exp(-|s_i|) \underbrace{\exp(-|s_i - s_{i+1}|)}_{g(s_i, s_{i+1})} \tag{4}$$

for $\boldsymbol{s}$. Here, Z is the normalization constant. Comparing Eq.(4) with the Laplace distribution, the modification is the multiplicative factor $g(s_i, s_{i+1})$. The function $g(s_i, s_{i+1})$ has a high value when s_i and s_{i+1} are correlated. As a result, $p(\boldsymbol{s})$ assigns a high probability for correlated sparse components and is a candidate for the estimation of our topography, which was defined using linear correlation. With this pdf, we can formulate the likelihood for the ICA model and obtain the following objective function L for the estimation of $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d)^t = \mathbf{A}^{-1}$,

$$L(\mathbf{W}) = J_1(\mathbf{W}) + J_2(\mathbf{W}), \tag{5}$$

$$J_1(\mathbf{W}) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{d} |\boldsymbol{w}_i^t \boldsymbol{x}(t)| + \log |\det \mathbf{W}|, \tag{6}$$

$$J_2(\mathbf{W}) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{d} |\boldsymbol{w}_i^t \boldsymbol{x}(t) - \boldsymbol{w}_{i+1}^t \boldsymbol{x}(t)|. \tag{7}$$

The vector $\boldsymbol{x}(t)$ denotes the $t-$th observation of the data, $t = 1, 2, \ldots, T$. Note that there are no constraints on $\mathbf{W}$.

3 Validation on artificial data

3.1 Experimental conditions

To validate that the objective in Eq.(5) is able to estimate a model with dependent sources as in Eq.(3), numerical experiments are performed on artificial data. First, $\boldsymbol{s}$ is sampled from Eq.(3) and the boundary of s_i is ring-like. Then, $\boldsymbol{x}$ is generated based on Eq.(1) where the entries in $\mathbf{A}$ are randomly generated. The dimension of the data is $d = 20$ and the number of samples is $T = 200000$. Preprocessing is performed by multiplying $\boldsymbol{x}$ with a whitening matrix $\mathbf{V}$. No dimensionality reduction was performed. Moreover, the absolute value $|a|$ is approximated as $\log \cosh(a)$. The initial values of $\mathbf{W}$ are sampled from the normal distribution. If our estimation is correct, $\mathbf{P} = \mathbf{WVA}$ should be diagonal, or, because of the ring-like boundary, a "shifted" diagonal matrix.

3.2 Results with a local optimization method

Here, we maximize the objective function L by the (nonlinear) conjugate gradient method of Rasmussen [7]. The covariance matrices of the sampled and estimated $\boldsymbol{s}$ are shown in Fig.1(a) and (b). Both matrices have a band matrix structure. Thus, our model can qualitatively correctly estimate the covariance matrix. However, $\mathbf{P}$ is dissimilar to a (shifted) diagonal matrix (Fig.2(a)). This means that the estimated $\boldsymbol{s}$ have a random order, and hence that the topography in the original $\boldsymbol{s}$ has not been recovered.

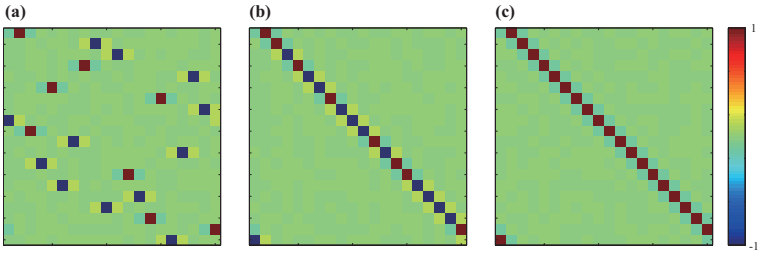

Figure 3: Matrices $\mathbf{P} = \mathbf{W}_{ortho}\mathbf{VA}$. In (a), raw $\mathbf{W}_{ortho}$ is used. For (b), $\mathbf{W}_{ortho}$ is permuted by hand and in (c), the signs are also changed by hand.

Table 1: Comparison of the values of the objective function corresponding to Fig.3(a-c).

	Fig.3(a)	Fig.3(b)	Fig.3(c)
$J(\mathbf{W}_{ortho})$	-15.9319	-15.9503	-15.8137
$J_1(\mathbf{W}_{ortho})$	-5.5700	-5.5700	-5.5700
$J_2(\mathbf{W}_{ortho})$	-10.3619	-10.3803	-10.2437

It is instructive to compare the obtained result with the result where the optimization is started at the true $\mathbf{W} = (\mathbf{VA})^{-1}$: After optimizing for this particular initialization, $\mathbf{P}$ was almost the identity matrix (result not shown). The values of the objective function for both initializations are

- $L(\mathbf{W}) = -14.7014$ for the random initialization (Fig.2(a)),
- $L(\mathbf{W}) = -14.5973$ for the true initialization.

Therefore, we can conclude that our result with the random initialization was only a local maximum. It looks like our optimization problem is much more difficult than, say, the one in TICA, and local maxima are a major problem that needs to be addressed.

3.3 New optimization method

We propose here an optimization method which is able to escape from the local maximum where the local gradient method above got stuck in. Empirically, we have found that orthogonalizing the $\mathbf{W}$ at the local maximum and then optimizing for the order and the signs of the $\boldsymbol{w}_i$ under orthogonality constraint is an effective means to escape from the local maximum. The relative contributions of the optimization for the order and the signs is shown in Table 1, together with Fig.3. Fig.3(a) shows $\mathbf{P} = \mathbf{W}_{ortho}\mathbf{VA}$, Fig.3(b) is the result when $\mathbf{W}_{ortho}$ is permuted by hand, and for Fig.3(c), the signs are also changed. From Table 1, we can see that $\mathbf{W}_{ortho}$ in Fig.3(c) gives the largest value of the objective function. This result clearly indicates that not only the order, but also the signs are important for our estimation. In addition, this result shows that $J_1(\mathbf{W}_{ortho})$ is insensitive to the change of the order and signs, which can be also seen from Eq.(6).

Mathematically, we can formulate this combinatorial optimization problem for the order and signs as follows:

$$\hat{\boldsymbol{k}}, \hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{k},\boldsymbol{c}} \underbrace{-\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{d} h(c_i \boldsymbol{w}_{k_i}^t \boldsymbol{x}_w(t), c_{i+1}\boldsymbol{w}_{k_{i+1}}^t \boldsymbol{x}_w(t))}_{J_2(\mathbf{W})}, \tag{8}$$

Here, $h(a,b) = \log\cosh(a-b)$, $\boldsymbol{k} = (k_1, \ldots, k_d)$ is an order vector where $k_i \in \{1, \ldots, d\}$ and $k_i \neq k_j$ for $j \neq i$, $\boldsymbol{c} = (c_1 \ldots c_d)$ is a sign vector where $c_i \in \{-1, 1\}$, and $\boldsymbol{x}_w(t)$ is the whitened data.

Combinatorial optimization problems consume in general much time and efficient methods are needed in practice. Fortunately, $J_2(\mathbf{W})$ has the remarkable property that it is a sum of functions of two variables only. This property suggests that we do not need to search all possible combinations and that the main problem can be divided into subproblems. Under this situation, dynamic

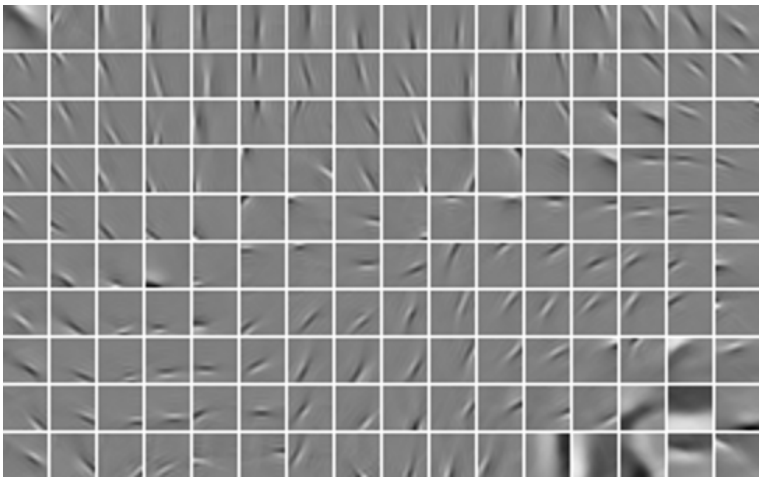

Figure 4: The learned topographic representation for natural images.

programming (DP) can efficiently solve our problem [8, 9]. We omit details because of the limited space.

As a whole, the flow of the optimization is as follows:

1. estimation of $\mathbf{W}$ by the conjugate gradient method as in section 3.2
2. orthogonalization and optimization of the order and signs by DP
3. re-estimation of $\mathbf{W}$ by using the optimized $\mathbf{W}_{ortho}$ from step 2 as the initial input to the conjugate gradient method

3.4 Results with the new optimization method

The result is shown in Fig.2(b): $\mathbf{P}$ is almost the (shifted) identity matrix. This result indicates that our estimation was performed correctly. Furthermore, the covariance matrix in the sampled $\boldsymbol{s}$ was qualitatively recovered after applying the new optimization (Fig.1(c)).

For comparison, the result from TICA[1] on the same artificial data is depicted in Fig.2(c) and (d). For this estimation, we found the same local maxima problem again (Fig.2(c)) , and thus, we ordered the components by hand so that the objective function of TICA is maximized (Fig.2(d)). At the maxima, TICA result produces a correct order, but the result is not as good as the one for the proposed model because the signs are not estimated correctly; in fact, they are arbitrary since TICA is invariant to the signs of the components. Furthermore, the cross-talk between the components is much larger than for the proposed model (Fig.2(b)).

4 Experiments on real data

4.1 Raw natural images

4.1.1 Experimental conditions

In this experiment, the data are 16×16 image patches extracted from natural images in the imageICA package.[2] The total number of patches is $T = 200000$. As preprocessing, the DC component of each patch is removed and then, whitening and dimensionality reduction are carried out by PCA. We retained 160 dimensions.

In this experiment, it is assumed that the components in the model are arranged on a two dimensional lattice and that component $s_{i,j}$ is correlated with the horizontal ($s_{i,j\pm1}$), vertical ($s_{i\pm1,j}$) and

[1] Matlab code for TICA is available at http://www.naturalimagestatistics.net/

[2] Available at http://www.cs.helsinki.fi/u/phoyer/software.html

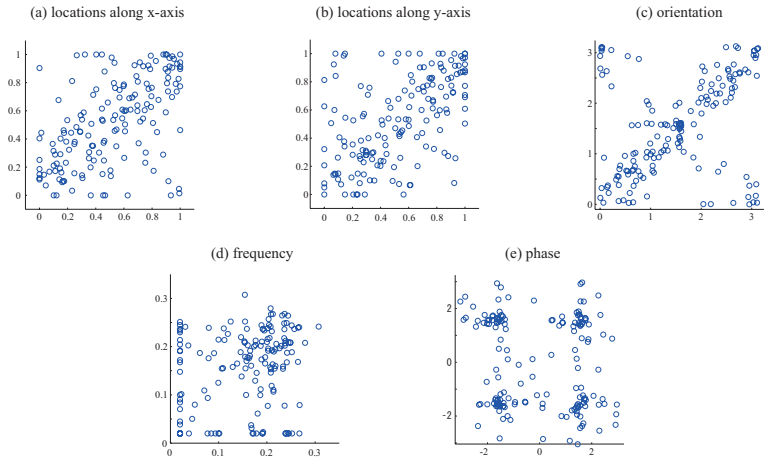

Figure 5: Scatter plots of Gabor parameters of adjacent basis vectors in Fig.4. The unit in (d) is cycles per pixel.

diagonal components ($s_{i\pm1,j\pm1}$ and $s_{i\pm1,j\mp1}$). The objective function and the optimization method for this two dimensional lattice is similar to the one for the one dimensional lattice, which we have presented above. Details will be described in a future article.

4.1.2 Results

The estimated basis vectors are presented in Fig.4. The basis vectors have localized, oriented and bandpass properties. These properties are qualitatively similar to those of the basis vectors for basic ICA [10, 11]. Adjacent vectors share properties such as spatial location and orientation. The whole topographic map resembles the one obtained by TICA.

To clarify the differences to TICA, we fitted Gabor functions to the basis vectors and compared the parameters with the TICA results. Scatter plots for the Gabor parameters of adjacent basis vectors are given in Fig.5. For the spatial location and orientation, adjacent basis vectors show clear correlation (Fig.5(a-c)). Frequency concentrates on high values (Fig.5(d)). These properties are much similar to TICA [12]. However, there is a strong difference for the phase (Fig.5(e)). The scatter plot for the phase in TICA is random, having no clear structure [12]. But Fig.5(e) shows that four clusters exist in our case. We estimated the center points of the four clusters by fitting Gaussian mixtures; interestingly, the center points are approximately $\pm\pi/2$. This result means that most of the basis vectors have an odd-type spatial distribution, *i.e.*, they represent edges instead of bars.

As a further comparison with TICA, we have investigated the co-linearity of adjacent basis vectors. Our findings are shown in Fig.6. The center point of each basis vector is projected on the z-axis, which is the preferred orientation of the adjacent basis vector, and the orthogonal z'-axis. Fig.6(b) is the scatter plot of the projected center points. For comparison, the same was done for the basis vectors from TICA. On the z-axis, the distribution of red circles (the proposed model) appears to be wider than the distribution of the blues ones (TICA). In fact, the variance for z for the proposed model is almost two times larger than for TICA, while the variance for z' is the same in both models (Table 2). Thus, adjacent basis vectors for the proposed model tend to be co-linearly arranged and more widely spaced along the preferred orientation. Such a property could be useful for the representation of long edges.

4.2 Complex cell outputs

Next, we applied the proposed model on the output of a fixed complex cell model when the inputs are natural images [13, 14].

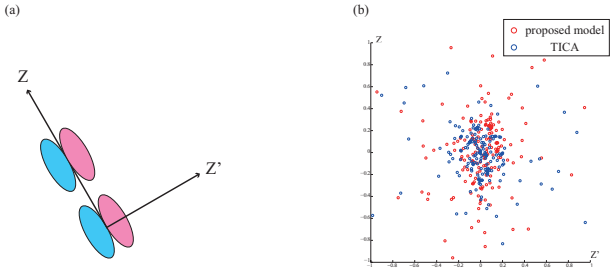

Figure 6: (a)A sketch of two adjacent co-linear basis vectors. z denotes the direction of preferred orientation and z' is the orthogonal direction. (b) The distribution of center points of the basis vectors using the $z - z'$ coordinates. Red and blue circles denote the center points of the proposed model and TICA, respectively.

Table 2: Comparison of the variance of z and z' for the proposed model and TICA.

	proposed model	TICA
var(z)	0.1129	0.0593
var(z')	0.0655	0.0673

4.2.1 Experimental conditions

The outputs of the complex cells are computed as

$$x'_k = \left(\sum_{x,y} W_k^o(x,y) I(x,y) \right)^2 + \left(\sum_{x,y} W_k^e(x,y) I(x,y) \right)^2, \tag{9}$$

$$x_k = \log(x'_k + 1.0), \tag{10}$$

where $I(x,y)$ is a natural image patch (size: 24×24), and $W_k^o(x,y)$ and $W_k^e(x,y)$ are odd- and even-symmetric Gabor receptive fields with the same parameters for spatial position, orientation and spatial frequency. In this experiment, the complex cells are arranged on a 6×6 spatial grid. For each position, there are cells with four different orientations and one frequency band. The total number of complex cells is $6 \times 6 \times 4 = 144$. To compute $\boldsymbol{x} = (x_1, x_2, \cdots, x_{144})$, we used the *contournet* matlab package.[3] As preprocessing, first, the DC component of $\boldsymbol{x}$ is removed and then, the variance of each component of $\boldsymbol{x}$ is standardized to one.

In this experiment, the estimation is performed on a one dimensional lattice. The procedure of the estimation is the same as in section 3.3.

4.2.2 Results

The estimated higher order basis had three prominent groups of features. In Fig.7(a), we show for each group a subset of the features. Adjacent basis vectors within each subset tend to have similar properties: the features in the upper figure show end-stopping, the features in the middle are star-like, and the features shown in the bottom figure are long contours. A further interesting point is that these three subsets are in the whole topographic map clearly separated from each other. Next, to verify that the features are not an artifact of the processing done with the complex cell model, we performed another experiment. In this experiment, $I(x,y)$ in Eq.(9) is sampled from a Gaussian distribution with mean $\mathbf{0}$ with the covariance matrix of the natural images used in the former experiment. The other experimental conditions are the same as before. The estimated higher order basis has also star-like features and features which are center-surround-like (Fig.7(b)). This last kind of feature is similar to the end-stopping features in Fig.7(a) but the "inhibition" is symmetric, which is not the case for the end-stopping features. Furthermore, long contours do not exist for the Gaussian data. Thus, the star-like features are mainly related to the complex cell model while the other features in Fig.7(a) are due to the statistics of natural images.

[3] Available at http://www.cs.helsinki.fi/u/phoyer/software.html

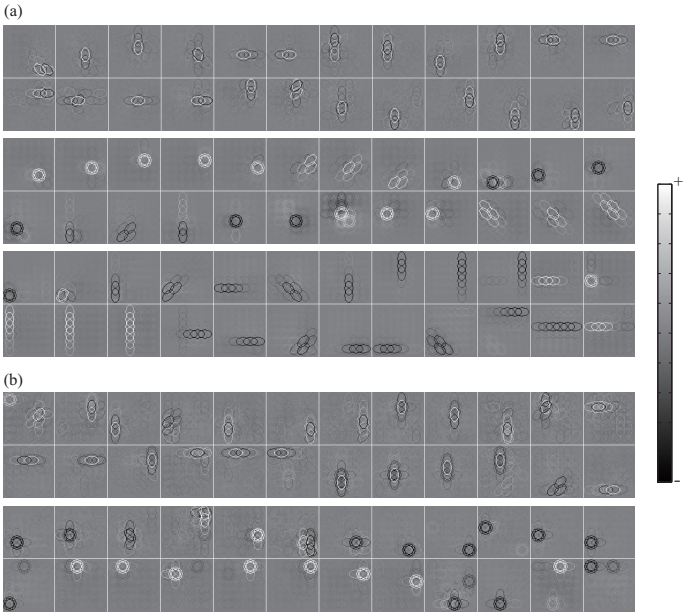

Figure 7: Estimated higher order basis by using (a) natural images and (b) noise inputs. In both cases, only a subset of the complete topographic map is shown.

5 Discussion and conclusion

In this paper, we proposed a new statistical model to estimate topographic representations. In the model, it is assumed that all components are sparse, and that adjacent components are linearly correlated while distant ones are as independent as possible. To estimate correlated sparse components, we derived a new optimization method based on dynamic programming since local gradient methods seemed to get stuck in quite bad local maxima.

The proposed model is related to topographic ICA, but there are fundamental differences. First, we model linear correlation, while TICA models energy correlation and constrains the components to be linearly uncorrelated. Second, the proposed model solves in a way the sign indeterminacy problem of ICA (or TICA) since linear correlation determines the sign of each component as a function of the adjacent components. Only the global sign of the matrix $\mathbf{A}$ remains undetermined.

When applied to natural images, the main difference between the model presented here and TICA is seen in the behavior of the phases (Fig.5(e)). The phases for basis vectors in the proposed model are approximately $\pm\pi/2$ and most of the basis vectors have an odd-type spatial distribution. The reason for this nonrandom phase property could be that presumably, the prominent feature of natural images is sharp step edges [15]. The second difference is the co-linear property of the basis vectors (Fig.6(b)). Neighboring basis vectors in the proposed model tend to have co-linear directions. This property could be due to the presence of long edges in natural images.

Using the outputs of complex cells, previous research on natural images has already discovered the emergence of long contour features by applying non-negative sparse coding [13] and ICA [14]. However, these results lacked a topographic representation. In addition, even though the adjacent basis vectors for the proposed model have similar properties, Fig.7 shows the emergence of a new kind of topographic representation.

Acknowledge

H. Sasaki was supported by Grand-in-Aid for JSPS Fellows. H. Shouno was partly supported by Grand-in-Aid for Scientific Research (C) 21500214 and on Innovative Areas, 21103008, MEXT, Japan. M.U.G. and A.H. were supported by the Centre-of-Excellence in Algorithmic Data Analysis. The authors wish to thank Shunji Satoh and Jun-ichiro Hirayama for their helpful discussion.

References

[1] A. Hyvärinen, J. Hurri, and P.O. Hoyer. *Natural Image Statistics: A probabilistic approach to early computational vision*, volume 39. Springer-Verlag New York Inc, 2009.

[2] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, 47(5):589–593, 2000.

[3] A. Hyvärinen, P.O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[4] S. Osindero, M. Welling, and G.E. Hinton. Topographic product models applied to natural scene statistics. *Neural Computation*, 18(2):381–414, 2006.

[5] Y. Karklin and M. S. Lewicki. A hierarchical Bayesian model for learning nonlinear statistical regularities in natural signals. *Neural Computation*, 17:397–423, 2005.

[6] G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J.L. Cantero. Measuring directional coupling between EEG sources. *Neuroimage*, 43(3):497–508, 2008.

[7] C.E. Rasmussen. Conjugate gradient algorithm, version 2006-09-08. 2006.

[8] R.E. Bellman. *Dynamic programming*. Princeton University Press, 1957.

[9] R.E. Bellman and S.E. Dreyfus. *Applied dynamic programming*. Princeton University Press, 1962.

[10] A.J. Bell and T.J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[11] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[12] A. Hyvärinen and P.O. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423, 2001.

[13] P.O. Hoyer and A. Hyvärinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 42(12):1593–1605, 2002.

[14] A. Hyvärinen, M. Gutmann, and P.O. Hoyer. Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in v 2. *BMC Neuroscience*, 6(1):12, 2005.

[15] L.D. Griffin, M. Lillholm, and M. Nielsen. Natural image profiles are most likely to be step edges. *Vision Research*, 44(4):407–421, 2004.